Developing Onomastic Gazetteers and Prosopographies for the Ancient World through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People

Yanne Broux[1], and Mark Depauw[2]

[1]Research Foundation - Flanders (FWO) / KU Leuven, Belgium
[2]KU Leuven, Belgium

Abstract. Developing prosopographies or onomastic lists in a non-digital environment used to be a painstaking and time-consuming exercise, involving manual labour by teams of researchers, often taking decades. For some scholarly disciplines from the ancient world this is still true, especially those studying non-alphabetical writing systems that lack a uniform transcription system, e.g. Demotic. But for many others, such as Greek and Latin, digital full text corpora in Unicode are now available, often even freely accessible. In this paper we illustrate, on the basis of Trismegistos, how data collection through Named Entity Recognition and visualization through Social Network Analysis have huge potential to speed up the creation of onomastic lists and the development of prosopographies.

Keywords: Named Entity Recognition, Graph Visualization, Ancient Prosopographies

0 Introduction

Developing prosopographies or onomastic lists in a non-digital environment used to be a painstaking and time-consuming exercise, involving manual labour by teams of researchers, often taking decades. For some scholarly disciplines from the ancient world this is still true, especially those studying non-alphabetical writing systems that lack a uniform transcription system, e.g. Demotic. But for many others, such as Greek and Latin, digital full text corpora in Unicode are now available, often even freely accessible. In this paper we illustrate, on the basis of Trismegistos (TM; www.trismegistos.org) [1], how data collection through Named Entity Recognition (NER) and visualization through Social Network Analysis (SNA) have huge potential to speed up the creation of onomastic lists and the development of prosopographies [2].

TM started out as a metadata database for sources from ancient Egypt, between 800 BC and AD 800, although at its roots lies a prosopography of Ptolemaic Egypt, the so-called Prosopographia Ptolemaica. Over the past years it has grown to a

interdisciplinary platform, encompassing several interrelated databases of not only texts and the people mentioned in them, but also place names, ancient authors, ancient archives, collections, and publications. TM is now expanding its geographical scope to the ancient world in general (currently counting 359,107 texts), and TM unique numeric identifiers for source documents (clean URIs such as www.trismegistos.org/text/1234), are now used not only in the papyrological world (e.g. papyri.info) but also in epigraphy (e.g. the Europeana EAGLE consortium). The eventual goal of TM is to provide unique identifiers for all texts from the ancient world, both published and unpublished. This means that TM increasingly wants to be a platform pointing to places where information can be found about all texts from antiquity, thus facilitating cross-cultural and cross-linguistic research.

1 Named Entity Recognition for Onomastic Gazetteers

NER was originally developed by computational linguists in the 1990s, but quickly spread to other fields, such as biology and genetics [3] and is now gaining momentum in the Digital Humanities [4]. The problem with NER-systems, however, is that techniques designed for one genre or field do not necessarily work for others, due to specific text properties (some follow strict writing constraints, e.g. scientific or news articles, while others, such as email or tweets, are more informal), or due to language-related grammatical and syntactical formats. With their diacritic marks, their sometimes fragmentary state, the case system of ancient Greek and Latin, and the for the Western World aberrant onomastic systems with tria nomina or fathers' names instead of family names, our documentation provides a real challenge for the automated collection of names.

1.1 Creating a multi-tiered onomastic gazetteer

In 2008 Bart Van Beek and Mark Depauw developed a database structure for the information on people occurring in the sources (TM People), and a NER procedure to extract references to the people in a Greek full text corpus [5]. The latter was made possible by the cooperation of the Papyrological Navigator, which just then released an Open Access Unicode version of the text of the roughly 50,000 papyri and ostraca from Egypt present in the Duke Database of Documentary Papyri. In early 2014 Mark Depauw developed a parallel system for Latin inscriptions

In each case the NER method was rule-based and relied on a gazetteer of personal names. Initially this consisted of a small set of some few thousand names from Ptolemaic Egypt. But of course many new names (fortunately easily recognizable through capitals) had to be added, and a strategy needed to be developed to cope with the multilingualism of the sources and the declensions of the inflected languages Greek and Latin. This resulted in the distinction of three layers of onomastic information, each with their own database: names, name variants, and declined name variants. The first database, NAM, currently has 34,094 entries, e.g. the Greek name Apollonios. Each of these names is connected to a set of transliterations and variants in all possible languages. As a rule, only very minor dialectal or orthographical

variation is allowed in the 'native' language (e.g. Ἀπολλώνιος and Ἀπολλλώνιος); most of the variants are created by renderings of a name in other languages, e.g. ȝpwlnys, ȝpwrnys or ȝpllʿns in Egyptian. In all there are 148,637 variants in the NAMVAR database. Finally, for each of the variants the various declined forms were created, to cope with that special type of variation: examples are Ἀπολλώνιου (genitive) or Ἀπολλώνιωι (dative). This NAMVARCASE database is the largest with 628,351 entries, and it is this set which was used as a gazetteer for the rule-based NER. This resulted in 510,533 attestations of the name variants, as tagged in the full text.

1.2 Distilling genealogical information from identifications

Building on the onomastic gazetteer, rules were then developed to cope with the combination of names, or more correctly declined name variants, in the identification of individuals. For the earlier texts in the Greek corpus, this was relatively uncomplicated, since in that onomastic system the standard way of identification is just a name followed by a father's name (in the genitival declined form). Already in the Ptolemaic period, however, there are complications with the use of double names, and in the Roman period not only are the names of more family members used (mother, paternal and maternal grandfather, ...), but also the Latin onomastic patterns are used more and more frequently. These imply the use of multiple names of different types (praenomen, nomen gentilicium, cognomen) for a single individual, as in Gaius Iulius Caesar.

To cope with this variation, for Greek a set of 164 rules was developed to interpret the clusters of onomastic identification. Criteria were the linguistic nature of the names (Latin names are not combined in the same ways as Greek or Egyptian ones), the case of the name (genitives being used to identify fathers), and the combination of the names with selected non-onomastic terms of identification, often referring to kinship (son, mother, ...). This allowed distillation of the genealogical information provided in the source. For the Latin inscriptions, a new start from scratch was made, because of the almost exclusive use of the Latin onomastic system and the very different composition of clusters, including also other types of elements such as the tribus (a geographical affiliation for Roman citizens).

1.3 Human intervention for quality control and intratextual identification

At this stage a human check was performed on the NER. This included tasks which were not so easy to automate: interpreting declined name variants as attestations of a specific case where the mere form was ambiguous; deciding whether some ambiguous entries were toponyms or anthroponyms; and reviewing the results of the cluster interpretation rules and adding relevant information where necessary.

All this could be labelled 'quality control', but we also decided to rely exclusively on humans for the logical next step when developing a prosopography, i.e. the identification of namesakes as attestations of the same person. Since the systematic review was performed text per text, only intratextual identifications were implemented.

2 Data visualization and network analysis to assist the creation of prosopographies

In its current state, TM People can thus not be called a prosopography, since the identification of namesake individuals is a crucial aspect of this type of scholarly tool. Nevertheless TM People has already proven its worth through quantitative analysis, using descriptive statistics to chart the reflection of social and religious changes in name giving in Greco-Roman Egypt [6-9]. Now we are taking things a step further by using data visualization and network analysis, both to optimize the database and to gain new insights into the social structures of ancient Egypt.

2.1 The problem of homonymy

Homonymy was fairly common in the ancient world. In village communities similar names were common. In families, names were often passed down every other generation, a way to express kinship in many societies where family names did not exist, as e.g. in Egypt and the Greek world. Moreover, for many individuals, we have not much more than their name: no titles or other status markers, no occupation, no "address", often not even an exact date for the text in which they are mentioned. All this makes it difficult to distinguish between one person and another, or, reversely, one becomes (too) cautious when matching attestations. These prosopographical identifications involve complex reasoning and can thus not easily be automated, but on the other hand our data set is simply too large and complex to review each attestation individually. For this reason we decided to adopt network visualization to facilitate the identification of people appearing in multiple texts [10].

2.2 Network visualization

With the help of network visualization, however, we are able to take into account an extra level: "communities", in this case in a rather abstract form, meaning people appearing together in different texts. Thanks to the interlock structure of the texts database (TEX) and the person attestation database (REF) in Trismegistos, a two-mode network of people-in-texts can easily be extracted and converted into a one-mode person-to-person network. This network can be checked swiftly for clusters of people reoccurring in several documents: these are most likely the same individuals. Visualizing our data in this manner presents us with a structured overview of the entire set, allowing us to achieve quicker results than when plodding through each individual record in the database.

A crucial element for the identification is of course the date of the document in which people are attested. Many ancient texts, however, do not mention a date, and in those that do, especially letters, the standard dating formula consists of the regnal year, followed by the month and the day. The name of the pharaoh (or king or emperor) is often omitted because it was obvious at the time of writing. In periods of unrest with contending rulers, perhaps even the year was left out because it allowed the scribe to remain impartial. In other cases, the part containing the date is damaged or lost. As a result, documents are often assigned to a broad span of time, e.g. 332-30

BC (= Ptolemaic period) on the basis of palaeography or content (a certain event, phrase, title, name, …). When exploring prosopographical identifications, for example, this can be particularly frustrating.

Again, data visualization can help us out here. When adding the dates of the texts in which the people appear as attributes to the same network generated above for the identification of individuals, broad date ranges can be narrowed down when these people are linked to others with a more accurate date. In a next stage, by combining the texts, the regnal years, and the people mentioned in the texts in one network, simple network concepts, such as the geodesic distance, can help to assign regnal years to a specific ruler when he is not mentioned in the text.

2.3 Letters from Elephantine

A group of Demotic letters from Elephantine, an island in the Nile on the border of Egypt and Nubia, serves as an excellent test case to illustrate the abovementioned methods. Trismegistos records some 146 letters from the fourth and third centuries BC, of which only 9 are (tentatively) dated to a specific year. The majority is simply attributed to 399-200 BC. Half of these texts mention a regnal year, but no ruler. We believe that a significant number should be assigned to the middle or the second half of the fourth century BC, around the time of Nectanebo II, the last pharaoh of the 30th dynasty (360-343 BC), and the Second Achaemenid period (343-332 BC).

These letters contain 450 attestations of individuals, but the identification of these people has only been carried out on a very limited scale, on the basis of the information given in the Prosopographia Ptolemaica [11]. By combining the two methods of personal identification and text dating, a pre-Ptolemaic date seems the most likely option for many of the texts, as we hope to show below.

Table 1. Excerpt of the nodelist of the one-mode Elephantine network

Id	Nodes	Nam_id	Date
8066	Eschnoumpmetis s. of E	175	303 BC
14501	Parates	752	216 BC
16284	Demetrios	2734	217 BC
16416	Es-onour-neb-shait	165	399-200 BC
16426	[ ]-sha-ti	1105	399-200 BC
16713	Esnebonychos	187	399-200 BC

Table 2. Excerpt of the edgelist of the Elephantine network

Source	Target
8066	16871
8066	17477
8066	57155
14501	57156
16284	16831
16284	16965

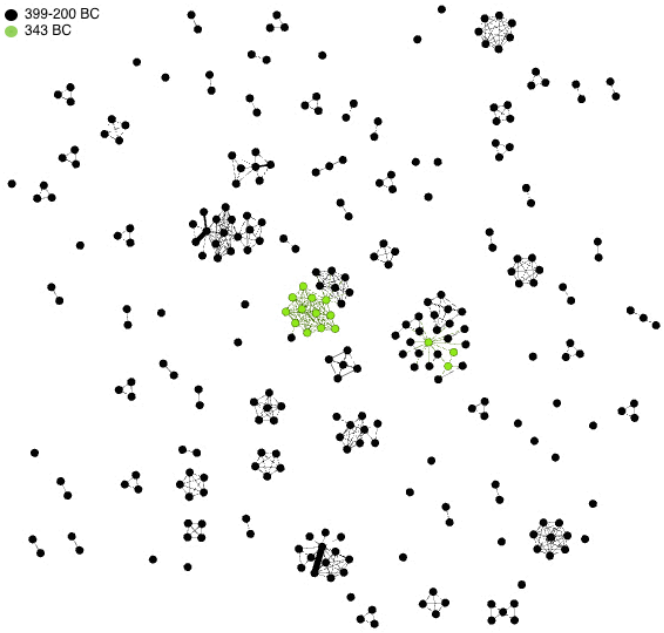

Fig. 1. Original Elephantine one-mode network

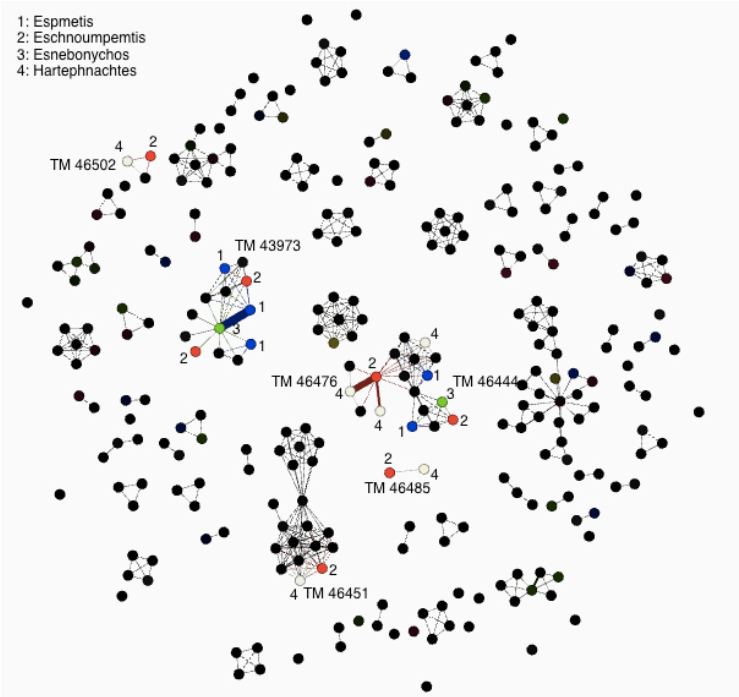

Fig. 2. Identifying individuals in the Elephantine network

People-in-texts to people-to-people. We started out with a two-mode affiliation matrix, listing all the people and the texts in which they are attested. This was converted into a one-mode network connecting those people mentioned together in one or more texts. Table 1 is a sample of the nodelist: it includes the person's unique numeric identifier[1], the person's name and patronymic, the name ID, and the date of the text in which he appears (if a person is attested in more than one text, the most accurate date was chosen). The edgelist consists of a simple adjacency matrix linking the person IDs (Table 2).

In this network, we discerned four clusters where precisely dated nodes (green = 343 BC) were combined with broad ones (black) (Figure 1). The dates of these blue nodes (= people), and consequently also the texts in which they appear, could therefore be narrowed down from 399-200 BC to 399-300 BC (green).

The next step was to check whether it was possible to identify any of the individuals, based on the reoccurrence of certain patterns of names (Figure 2). When highlighting the four most common names, several combinations appeared in six different texts: Eschnoumpmetis and Hartephnachtes are mentioned together in four (TM 46451, 46476, 46485 and 46502), the first of which is assigned to 343 BC; Espmetis (x2) and Esnebonychos in one (TM 43973); and Eschnoumpmetis, Hartephnachtes and Espmetis in another (TM 46444). Twice, both Eschnoumpmetis and Hartephnachtes are identified as sons of Esnebonychos (TM 46444 and 46476). In TM 43973, one of the men called Espmetis is a son of Esnebonychos as well: most likely the same Espmetis mentioned together with Eschnoumpmetis and Hartephnachtes before (TM 46444). Finally, in TM 43973 we also have an Espmetis son of Es-pa-nty-hut-neter and an Esnebonychos son of Es-pa-nty-hut-neter, perhaps the same Esnebonychos who is listed as the father of Eschnoumpmetis, Hartephnachtes and Espmetis? If these identifications are correct, we can reconstruct the following family tree (Figure 3):

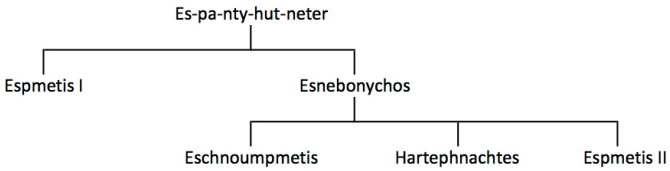

Fig. 3. Family tree of Eschnoumpmetis

After carrying out these identifications, we reconstructed a new network, and this time a giant component connecting all the green nodes, and several new blue ones, emerged (Figure 4). Again, the date of those blue nodes could be narrowed down to

[1] Since Trismegistos uses unique identifiers for people (www.trismegistos.org/person/1234), attestations (www.trismegistos.org/ref/1234), texts (www.trismegistos.org/text/1234) and names (www.trismegistos.org/name/1234), we use these instead of the actual names or publications to avoid confusion and spelling mistakes.

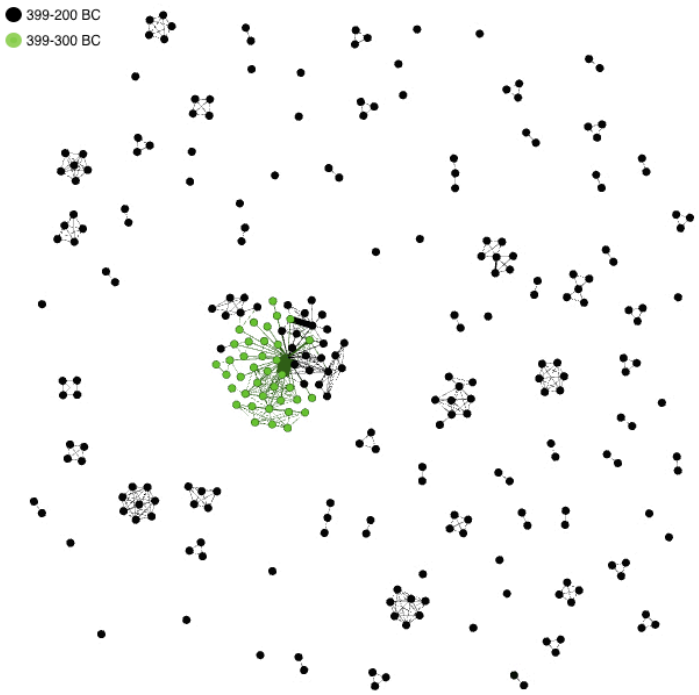

Fig. 4. New Elephantine one-mode network

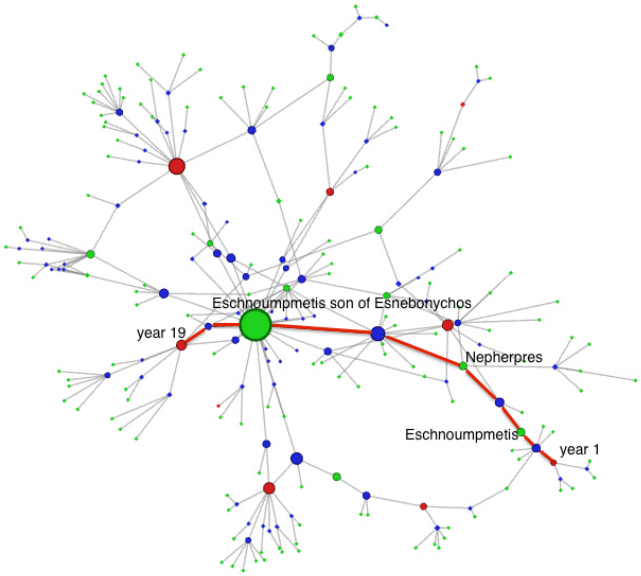

Fig. 5. Elephantine three-mode network

the fourth century BC. Some extra identifications could also be performed. In this cluster, the name Osoroeris (son of Teos) appeared four times, as well as in one of the unconnected components: they were identified as one and the same person, as well as three attestations of Nepherpres, who is always mentioned in texts together with Eschnoumpmetis (son of Esnebonychos). Finally, three nodes labelled Eschnoumpmetis son of Psammetichos (of which one in the giant component) were also merged.

All-in-one network. In a next stage, we gathered all the new prosopographical information, and constructed a three-mode network of people that appear in texts that in turn are linked to a regnal year (Figure 5). Our aim was to see if we could link year 19, which we believe to be the last year of Nectanebo II's reign, to year 1, the first of Artaxerxes III; they would correspond to 343 BC.

The shortest path between the two (red line), or geodesic distance, measures 8 in this case: year 19 – TM 46477 – Eschnoumpmetis son of Esnebonychos – TM 46451 – Nepherpres (or year 18) – TM 46615 – Eschnoumpmetis – TM 46443 – year 1. If we could identify the second Eschnoumpmetis with Eschnoumpmetis son of Esnebonychos, we would even get there in four hops, the absolute minimum to get from one year to another in this network. Unfortunately, there are at least two other people called Eschnoumpmetis (a son of Chnoum-machis and a son of Psammetichos), so this identification is far from certain. An alternative eight-hop route is year 19 – TM 46499 – P-oudja-metoues son of Psentaes – TM 46539 – year 18 – TM 46615 – Eschnoumpmetis – TM 46443 – year 1.

3 Conclusion

Named Entity Recognition and graph visualization have thus already proven to be tools that greatly facilitate the creation of new onomastic gazetteers and prosopographies. Yet there is substantial scope for improvement and further assistance of digital tools. In the current process, time-consuming human intervention remains indispensible at several stages. Also, the databases, NER-procedures, and graph visualizations and manipulations remain locked away in separate programmes. Further integration of e.g. database identification in the graph visualization or network-based automated suggestions for identification remain interesting prospects for the future.

References

1. Similar methods are applied in e.g. Klein, L.F.: The Image of Absence: Archival Silence, Data Visualization, and James Hemmings. American Literature 85:4, pp. 661-688 (2013)
2. Depauw, M., Gheldof, T.: Trismegistos. An Interdisciplinary Platform for Ancient World Text and Related Information. In: Ł. Bolikowski, V. Casarosa, P. Goodale, N. Houssos, P. Manghi, J. Schirrwagen (eds.), Theory and Practice of Digital Libraries - TPDL 2013 Selected Workshops. CCIS, vol. 416, pp. 40-52. Springer, Heidelberg (2014)

3. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Linguisticae Investigationes 30:1, pp. 3-26 (2007)
4. van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections. Literary and Linguistics Computing. The Journal of Digital Scholarship in the Humanities, (2014) (forthcoming)
5. Depauw, M., Van Beek, B.: People in Greek Documentary Papyri. First Results of a Research Project. Journal of Juristic Papyrology 39, pp. 31-47 (2009)
6. Broux, Y.: Double Names and Elite Strategy in Roman Egypt (Studia Hellenistica 54). Peeters, Leuven (2014)
7. Coussement, S.: Because I am Greek: Polyonymy as an Expression of Ethnicity in Ptolemaic Egypt (Studia Hellenistica 55). Peeters, Leuven (forthcoming)
8. Depauw, M., Clarysse, W.: How Christian was Fourth-Century Egypt? Onomastic Perspectives on Conversion. Vigilae Christianae: A Review of Early Christian Life and Language 67, pp. 407-435 (2013)
9. Jennes, G.: Inspired by the Gods: Theophoric Names in the Late and Graeco-Roman Periods in Egypt. Unpublished PhD dissertation, Leuven (2012)
10. Our approach leans toward the methods applied by Rossi et al. to a large database of French notarial acts from the 13th-18th centuries: Rossi, F., Villa-Vialaneix, N., Hautefeullie, F.: Exploration of a Large Database of French Notarial Acts with Social Network Methods. Digital Medievalist 9 (2013) (http://www.digitalmedievalist.org/journal/9/villavialaneix/)
11. *Prosopographia Ptolemaica* (Studia Hellenistica), 10 vols. Peeters, Leuven (1950-2002)